

修 士 論 文 の 和 文 要 旨

大学院 情報システム学研究科		博士前期課程	情報システム設計学専攻
氏 名	齋藤 裕明		学籍番号 0550017
論 文 題 目	木データのラベル類似性に着目した分類及び検索手法に関する研究		
<p>要 旨</p> <p>木は単純でありながら高い表現能力を持ったデータ構造であり、ノードにラベルが付いたラベル付き木は、半構造データや遺伝子情報など多様なオブジェクト表現に用いられている。</p> <p>これらの木として記述されたオブジェクトに対してパターン認識や分類、情報検索を行う際に木間の類似度を求める技術が重要になる。木間の類似度としては、2つの木をノードの挿入、削除、置換によって一致させる際の最小編集コストを類似度とする木編集距離(tree edit distance)が知られている。</p> <p>データの分類や検索を行う際の多くの目的は、そのデータの種類など、データの内容の類似性によって分類、検索を行うというものである。木編集距離では2つの木を一致させる際の編集コストを木間の類似度とする為、木データ間の内容を表すノードラベルの類似性と、ノード数や形状といった構造的な類似性の両方を同時に含んでしまう。その為、ノード数や形状が大きく異なる木データ間の木編集距離は、構造の違いを大きく反映し、ノードラベルの類似性が反映されないという問題点がある。</p> <p>そこで本論文では、2つの木を一致させる編集操作を分解し、ノードラベルを一致する編集操作から木データ間の内容的な類似性を表すラベル非類似度を提案する。ラベル非類似度は、2つの木間で同じラベルのノードが同じ位置にある割合を表している。また、ラベル非類似度は木の包含関係であるTree Inclusionを一般化したものとみなす事が出来る。</p> <p>木編集距離ではアプリケーション毎に適切なノードの挿入、削除、置換操作コスト(ノード編集コスト)を事前に設定する事が要求される。これは適用するアプリケーションに關しての専門的な知識が必要となる事を意味するが、このような専門知識を用意するのは難しいので、ノード編集コストを全て等しく設定する単一コスト木編集距離が一般的に用いられる。ラベル非類似度もノード編集コストを事前に設定する必要が無く、単一コスト木編集距離と同様に高い汎用性をもつ。更に、木編集距離の問題点であった構造の類似性による影響を受けない為、単一コスト木編集距離よりもユーザの分類、検索目的を満足する結果を得ることが出来る。</p> <p>ノイズを含む部分木の判別問題、半構造データの分類の2つの応用例にラベル非類似度を適用して有効性を検証した。ノイズ付き部分木判別では既存手法よりも高い判別精度と計算時間が早い事を確認した。半構造データの分類において、構造の類似性を表した構造非類似度とラベル非類似度を適切に組み合わせる事により、ユーザの目的を適切に反映した分類結果が得られる事を確認した。</p>			